ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Graduate Education in Statistics and Data Science: The Why, When, Where, Who, and What

Marc Aerts,[1] Geert Molenberghs,[1,2] and Olivier Thas[1,3,4]

[1]Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, BE3590 Hasselt, Belgium; email: marc.aerts@uhasselt.be

[2]Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), KU Leuven, 3000 Leuven, Belgium

[3]National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Keiraville, New South Wales 2500, Australia

[4]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, 9000 Ghent, Belgium

Annu. Rev. Stat. Appl. 2021. 8:25–39

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040620-032820

Keywords

computer science, curriculum, data science, education, graduate, interdisciplinary, statistics

Abstract

Organizing a graduate program in statistics and data science raises many questions, offering a variety of opportunities while presenting a multitude of choices. The call for graduate programs in statistics and data science is overwhelming. How does it align with other (future) study programs at the secondary and postsecondary levels? What could or should be the natural home for data science in academia? Who meets the entry criteria, and who does not? Which strategic choices inevitably play a prominent role when developing a curriculum? We share our views on the why, when, where, who and what.

1. WHY?

Successive waves of varying degrees of innovation have run through the statistics community during the past 75 years: epidemiology and nonlinear models such as logistic regression, Bayesian model fitting, complex multivariate and hierarchical models, smoothing methods, bootstrap, data mining, bioinformatics, omics, big data, and so on. Not surprisingly, they were often preceded by technological breakthroughs in computer science and software development and/or by innovation in one or more substantive sciences. With varying interinnovation time lags, these waves have rippled through all aspects of our work, including in education in statistics, at all levels from secondary school up to postgraduate university programs.

Although the current data science wave has similarities with earlier waves, especially the bioinformatics and big data waves, the data science wave is occurring on a much larger scale. It was initiated not only by scientific or intellectual developments but also by societal, political, and commercial developments. More than ever, detailed and personalized data are available (and not always with recognized consent of the individual), and more than ever, such data imply knowledge, insights, and evidence-based decisions and treatment choices to best benefit the patient or client. But other commercial interests may have a dominant impact on how data science will develop further. The current data science hype has many characteristics in common with bioinformatics, including the availability and integration of several big (public and private) data resources. There is, of course, lack of clarity and even controversy about the definition of data science. One's definition has become, in itself, a discrete random variable with probabilities highly depending on characteristics of the user's profile (e.g., engineer, computer scientist, theoretical physicist, statistician). Bioinformatics was, at the early stages, often claimed by computational biologists and/or computer scientists, starting with the flawed linguistic observation that the term "bioinformatics" itself was an indication that informatics, rather than statistics, played the center-stage role. Meanwhile, and for over a decade, it has been recognized that statistical methodology is a key component of bioinformatics, and statisticians started to use the term "statistical bioinformatics" to label this component more explicitly. One could think of the data science movement as a final, all-encompassing stage in the expanding process that started with bioinformatics in life sciences and then jumped to omics data and to big data in any other application, especially marketing, business, and related fields, and further to the social and political sciences.

We do not aim to present yet another definition of data science. Definitions vary from exhaustive lists of subjects of knowledge and skills to just defining data science as the science of learning from data, with everything that this entails (e.g., Donoho 2017), the latter covering knowledge and skills of data management, data wrangling, data visualization, planning of studies, statistical methods, models, inference, artificial intelligence, machine learning, and so on. It naturally also includes sufficient knowledge about the substantive area within which an application or methodological development is situated. For example, cameras recognizing individual behavior and translating these data into actions may be perceived as data science.

The wide scientific aura around a data scientist might explain the tendency to identify such a data scientist as an engineer, who typically has enjoyed exposure to many scientific fields in an applied fashion during their education, as compared with, for example, the graduate of a primarily mathematical statistics program. We believe that there is room and a need for several varieties of data scientists, varying in their particular knowledge and skill sets but with sufficiently overlapping subsets. All are needed in our complex contemporary society.

In Section 2, we discuss at which stage in an individual's education exposure should be given to data science. Where such exposure and education should be administered is the topic of Section 3. Who holds responsibility over it and should be involved is tackled in Section 4, and

what a curriculum in data science at the graduate level should look like is approached in Section 5. An outlook to the future is offered in Section 6.

2. WHEN?

At young ages, children can already observe, discover, and experience basic concepts of data science in their everyday environments, including the quantitative and qualitative data omnipresent in our society, basic practices in collecting and sharing data, basic use of information and evidence from data, and the sensitivity of certain data. But little information seems to be available on educational strategies implementing a systematic progression of data science knowledge and skills across curricula, covering the age range of, say, 5–18 years. Heinemann et al. (2018) developed a data science curriculum for German secondary schools, designed as an interdisciplinary approach between mathematics and computer science education, with a strong focus on the societal aspects. An interesting and inspiring report entitled "The Integration of Data Science in the Primary and Secondary Curriculum," addressed to the Royal Society Advisory Committee on Mathematics Education (Pittard 2018), reviews Great Britain's primary and secondary national curricula (key stages 1–4), identifies the extent to which elements of data science exist within data-rich subjects, and identifies specific topics that are conducive to nurturing data science skills. The main challenges, barriers, and strategies to overcome these, as well as further opportunities, are discussed. We can only recommend that other governmental authorities worldwide conduct similar insightful exercises.

Data science education's interdisciplinary nature, extensive and complex prerequisites, numerous application areas, and rapid scientific development, as well as the huge demand for data scientists in industry and data science's growing impact on the future, make it a very important but challenging field for students, parents, teachers, curriculum developers, schools, universities, scientific societies, and governments. It is unquestionably one of the great challenges in contemporary education, but it also offers unique opportunities. These challenges cannot be faced, and these opportunities cannot be taken, by individual teachers or single organizations or authorities only, and the scientific and educational communities can hope for, but not wait for, appropriate measures and guidelines from governmental bodies, as these will be, in all likelihood, ill-timed, ill-conditioned, and with too limited an outlook. All initiatives are very much welcome, but as data science is intrinsically interdisciplinary, the approach for data science education to face the challenges and grasp the opportunities should be "inter-" in many facets as well: interdisciplinary, intercurricular, intergrade, and so on, around intercollaborative initiatives across study programs of different age groups, with involvement of secondary school teachers, professors, and so on. Data science education will face the same shortage of skilled educators at the primary and secondary levels that statistics education has been confronting. Indeed, this is a pervasive problem across disciplines and countries. For the rapidly evolving field of data science, which currently lacks a distinct place and list of end competencies in the primary and secondary curricula, a solution could be to support and facilitate collaboration between schools and universities. Here, universities could play a leading role, alongside offering bachelor's and master's programs in statistics and data science. Support could come from grants in innovative educational projects awarded by universities, government, and industry. There is a need for education policy and curricular choices in schools to educate students in data science, similar to or as part of the STEM (science, technology, engineering and mathematics) initiative (see, e.g., Hallinen 2019). A diverse palette of organized learning activities could stimulate and excite youngsters' interest in data science. This could include professors participating in meet-the-professor events in secondary schools. Senior students taking a master of statistics and data science could be involved in teaching data science in secondary school

(and getting credit for it in their own study program), and student-driven science fair projects and citizen science could be part of this initiative (Koomen et al. 2018, for example, provides an interesting account of middle school science fair projects inspired by citizen science monitoring). Of course, these activities should not just occur as fragmented and isolated learning units; rather, they should fit into a strategic plan with a growth-curve pattern for interested students.

Turning to undergraduate programs in statistics and data science, the 2015 special issue of *The American Statistician* (Horton & Hardin 2015) on statistics and the undergraduate curriculum includes several contributions on data science in statistics curricula, providing several examples and resources for instructors to implement data science in their own statistics curricula. De Veaux et al. (2017) proposed detailed guidelines for revising a major in data science, starting with redesigning existing courses in computer science, statistics, and mathematics, and guided by the relation of data science to the other sciences. In the next stage, they suggest transitioning existing courses further and developing new, more fully integrated courses, taking advantage of the efficiencies and synergies that an integrated approach to data science would provide. They propose a case-based and hands-on approach, as is common in fields such as engineering and computer science. We would like to add the field of applied statistics to this. We can only fully agree with the statement that the two pillars, computational (algorithmic, predictive) and statistical (inferential) thinking, should not be kept separate but rather should be integrated. Yavuz & Ward (2020) discuss the implementation of an introductory data science course, offered during a student's sophomore undergraduate training in statistics. It implements many of the best practices espoused by De Veaux et al. (2017), including interaction and a strong focus on teamwork.

De Veaux et al. (2017) mentioned in their report that the website **http://datascience.community/colleges** listed, at the time, 530 programs in data science, analytics, and related fields at more than 200 universities around the world. At the time of writing, the website now lists 618 programs, with a vast majority of these being master's degree and certificate programs offered both traditionally and online.

Since December 2016, the US National Academy of Sciences has been holding a series of roundtables on postsecondary data science education. Each of four meetings per year focuses on a topic related to data science education or practice. These roundtables bring together representatives from academic data science programs, funding agencies, professional societies, foundations, and industry to discuss the community's needs, best practices, and ways to move forward. The objective is to help affected communities develop a coherent and shared view of the emerging field of data science and of how best to prepare large numbers of professionals to help realize the potential of this field. The written summaries are available at **https://nas.edu/dsert** and are a rich source of information based on the views of several experts with different backgrounds, experiences, and interests. Topics include the development of data science curricula and programs at two-year colleges, and its importance in the development of a diverse and inclusive workforce (one-third of the total undergraduate student population in the United States is enrolled at two-year colleges, citing Nicholas Horton from the minutes of meeting #11); promotion of data science for socially desirable outcomes; content, organization and alternative structures of data science PhD programs; and many others. On the European side, there is the European Data Science Academy (EDSA), funded by the Horizon 2020 Framework Programme of the European Union (**http://edsa-project.eu/**). Mikroyannidis et al. (2018) present this European initiative for bridging the data science skills gap across Europe and training a new generation of world-leading data scientists. The EDSA project has established a rigorous process and a set of best practices for the production and delivery of curricula and courseware for data science, as well as linking demand with supply.

So, on the question of when particular data sciences learning units on statistics and data science should be offered, we would like to respond: "at any age, starting at early age, and continuing lifelong." Not everyone will agree with this point of view. One could, for example, argue that it is better to wait until a student has been confronted with the urgency to learn statistics and data science when trying to analyze their own data (e.g., from a bachelor's thesis). This genuine interest in data, however, can also be obtained for children in secondary school by connecting with real-life situations and active citizenship. For instance, the coronavirus disease 2019 (COVID-19) crisis of 2020 offers endless opportunities to connect with data—including data on social contacts illustrating the importance of measures to change human contact behavior, and numbers of infected, hospitalized, and deceased individuals—and to discuss sensitive issues around such data, compare curves, and discuss time trends shown daily on the Internet.

Also, continuing education and lifelong learning become more important as the pace of scientific and technological innovation keeps on accelerating. Education in statistics and data science needs a holistic approach, maintaining and upgrading knowledge and skills in expanding contexts and environments. Therefore, it is important to teach students at the bachelor's and master's levels skills for self-learning, self-motivation, self-monitoring, self-adjustment, and self-sustainability—skills we come back to in Section 5.

In our view, a full two-year study program of, say, 120 study credits on statistics and data science is ideally offered as a master's program. We also think that initiatives across different study programs and age groups, which trigger the general interest of students at a young age, stimulate collaboration across borders, and fill gaps of expertise in statistics and data science at particular levels (e.g., by master's and PhD students, postdoctoral researchers, and professors in the field of statistics and data science contributing to education at the secondary level), can only be welcomed and should be supported in various ways. Industry can and should also take the opportunity to play its role [e.g., data scientists in industry acting as guest professors in master's programs or external supervisors of master's theses, or supporting and participating in so-called hack weeks (Huppenkothen et al. 2018)]. Industry and higher education must join forces to answer the large and growing demand for statisticians and data scientists in the coming years. Whereas statistics has too often been branded as boring, data science has been called the "sexiest job of the twenty-first century" (Davenport & Patil 2012) and is better positioned to capture the interest of young people. This momentum should be taken advantage of; all data scientists should consider contributing to education in data science.

3. WHERE?

Hicks & Irizarry (2018) argue that a statistics department that embraces applied statistics is a natural home for data science in academia. While statistics is not the only field contributing to data science, and while arguably there is no uniform best organizational format, there are several reasons why the statistics discipline might beneficially be placed at the center of organizing data science. Almost invariably, successful applied statistics groups, whether focused on a single or various substantive areas, bring under a single roof research, teaching, and consulting. When both methodological research and collaborative research with other fields of study are undertaken, such a department is in a good position to ensure sufficient breadth and depth of the data science curriculum. A group's focus in research and consulting should ideally translate into the signature features of the curriculum offered. For example, a group focusing primarily on biostatistics may be well suited to lead a graduate program in data science for biostatistics, but it may be challenging for it to support a program focused on, for example, engineering or economics. This is important for master's-level education, and even more so for PhD-level training, to ensure that PhD candidates are able to work on relevant and challenging projects of sufficient breadth.

There should, therefore, be close collaborations between statisticians and data scientists on the one hand, and the substantive areas covered in the graduate program on the other hand. Hicks & Irizarry (2018) argue that applications should be brought much more to the forefront than is currently the case. The problem should come first, and the teaching should be linked to one or a number of substantive problems, even when more theoretical and abstract concepts are being taught. It is possible that a given institution would not encompass all the specialties envisaged. Wherever possible, it is advisable to consider interinstitutional collaborative efforts. This is relatively easy in geographical areas where several complementary institutions are available within a relatively short distance. For example, a university with a large engineering school could collaborate with an institution that is known for its medical research. Fortunately, because of the increasing availability of electronic and distance learning tools, the physical distance between institutions is becoming less and less of a limiting factor.

The above makes clear that a signature feature of applied statisticians is the habit of collaborating within their own field with colleagues with different interests on the one hand, and with colleagues from the substantive fields on the other. This is a vital asset because data science requires more, not less, collaboration to give curricula the broad and high-quality basis they deserve. In other words, the collaborative network is expanding thanks to data science, incorporating especially engineering, numerical mathematics, and computer science groups, and teams from other application areas, leading to more so-called use cases. In some institutions, often larger or older ones, there may not be a single applied statistics group, but various statistics groups may coexist, with each group specializing in a particular area of application. In others, often smaller or younger ones, statisticians working with various substantive areas may find themselves in the same organizational unit. In both cases, it is perfectly possible to organize a data science graduate curriculum. Hasselt University in Belgium is a smaller, younger university with a relatively large center for statistics. Their research, consulting, and graduate teaching all focus on biostatistics, epidemiology and public health methodology, and bioinformatics, which are then quite naturally the themes within the Master of Statistics and Data Science program. A data science institute at the same university encompasses this structure, as well as researchers from computer science, medicine, and economics. Leading data science institutes exist at Columbia University (encompassing computer science, statistics, engineering, and operations research) and at the University of Virginia (with computer science, statistics, and systems engineering departments participating). At the large and older KU Leuven, statisticians are located across ten different faculties, but their paths cross in the university-wide Leuven Statistics Research Centre, which in turn is the organizational structure for the Master of Statistics and Data Science program. Indeed, when statisticians are spread over various entities, it is important to have efficient lines of communication between them and, ideally, a superstructure such as a university-wide center or institute for statistics (and data science). At Harvard University, the Institute for Applied Computational Science groups statistics, computer science, and applied mathematics. In some places, a data science program is organized jointly by various quantitative departments, but there is no dedicated data science center or institute: At Kennesaw State University, statistics, computer science, and mathematics all contribute; at the University of Colorado, it is organized by biostatistics, informatics, computer science, and engineering; biostatistics and medical informatics are the contributing fields at the University of Wisconsin-Madison; at the University of British Columbia, statistics and computer science co-organize the program; and Tufts University's program is a collaboration between computer science and engineering as the basis for their data science master's—an example where statistics is not formally part of the organizational structure. At the University of Kansas, a data science program is organized by the biostatistics department, and to this end, statisticians with computational background are recruited to the faculty. At the University of Toronto, the

Master of Applied Computing program is led by the computer science department, with the data science concentration in the hands of the statistics department. In various other places, there is a partnership between statistics and a substantive area. For example, collaborations between statistics and mathematics on the one hand, and the business or management school on the other, exist at Texas A&M, Clemson University, and the Massachusetts Institute of Technology.

The variety of data science program structures underscores that there is no single best way of organizing a data science graduate education program. But the institutional environment should ensure that the recursive data cycle (De Veaux et al. 2017), made up of obtaining, wrangling, curating, managing, and processing data, is steeped in practice. Likewise, Wild & Pfannkuch (1999) refer to this as the "problem, plan, data, analysis, conclusion" cycle. No matter what the institutional context is, the more researchers from various disciplines are integrated and feel like part of a larger entity, the higher the chance that the cycle will be effective.

Because of the interdisciplinary and multidisciplinary nature and the extent of the collaborative network needed, it is futile to assume that establishing a single data science department, rigidly existing next to other departments, is a fruitful approach. Rather, because of the multidisciplinary nature, efficient networking-type structures are needed, of which academic personnel can be a member without having to give up membership in their conventional department. Kane (2014) states that statisticians are situated at the confluence of statistics, computer science, and the substantive area. Cleveland (2014) itemizes the various disciplines and fields that should be present in data science research, which mirrors the needs in the accompanying graduate programs, of course. They are multidisciplinary investigation, models and methods for data, computing with data, pedagogy, tool evaluation, and theoretical foundations of data science.

Hicks & Irizarry (2018) refer to type A and type B data scientists. Type A is concerned more with the statistical and methodological side, in view of answering real-world questions; Type B refers to the coding side of data science and is rooted in engineering and computer science. In a way, the aforementioned Tufts University example is of Type B. Whereas Hicks & Irizarry (2018) focus in their guide to teaching data science on Type A, arguably, at graduate level, it is important that both types are brought together in a flexible, perhaps virtual structure that facilitates interaction and collaboration.

Evolutions of this type are not new to statisticians, even though it may seem so in the current day and age. One of the earlier similar events was the advent of epidemiology. To this day, there are a variety of ways in which statisticians and epidemiologists have organized themselves around each other. It is telling that in many institutions, scholars belong to both. Similar evolutions took place with the coming of, respectively, statistical genetics, bioinformatics, the omics, and big data.

Broadly, the drivers of new evolutions are either a quantum leap in a substantive field (e.g., the decoding of the genome), necessitating new or expanded quantitative methodology; a major step forward in quantitative sciences (e.g., increasing computational power, data capture from social media); or both (wearable medical devices). One often drives the other. For example, the advent of generalized linear models, such as logistic regression, which is very commonly used in epidemiology, was converted to practical use thanks to advances in computer hard- and software. In this sense, Cleveland's (2014, p. 417) broad view on data science is important: "A very limited view of data science is that it is practiced by statisticians. The wide view is that data science is practiced by statisticians and subject matter analysts alike, blurring exactly who is and who is not a statistician."

We can learn from these earlier (r)evolutions in many ways. For example, programs in epidemiology and public health may be stand-alone or organized jointly with (bio)statistics programs. The same is true for bioinformatics. Various organizational forms are possible, and here also, there is no single best format. For example, at Hasselt University, the Master of Statistics and Data

Science program encompasses tracks in biostatistics, quantitative epidemiology, bioinformatics, and data science. At KU Leuven, the Master of Statistics and Data Science program organizes data science in a variety of tracks related to various fields of study (biometrics; social, behavioral, and educational sciences; industrial applications; business; official statistics; and theory). The bioinformatics curriculum at KU Leuven, however, is stand-alone, with some courses in common with the master's of statistics and interchange of faculty with other departments.

It seems evident, in this day and age, that we need to make our programs resilient. So, they should not rely purely on imparting knowledge in a face-to-face way. Rather, they should be interactive and involve two-way traffic between students and instructors. The use of distance learning and flipped classroom techniques should be part of a program in a routine fashion. As demonstrated by the 2020 COVID-19 crisis, it is of the essence that a program can revert to a fully online version by merely flipping a switch. High-quality programs may make use of massive open online courses, or at least small private online courses, to this effect.

4. WHO?

Flexibility is needed to accommodate various types of prior education at the undergraduate level and, for some students, also prior graduate education. It is now common that many students choose graduate education, or further graduation education, at institutions different from the ones where they obtained earlier degrees. Careful assessment of their prior knowledge and skills is necessary, not only to decide on admission but also to gauge whether a reduced or tailor-made curriculum is advisable. We can think of three types of students admitted to a data science graduate program: (*a*) those with undergraduate training in data science; (*b*) those with undergraduate training in one or a few of the contributing fields, i.e., statistics, engineering, computer science, or mathematics; and (*c*) those with undergraduate training in a sufficiently quantitative subject area. All of this implies that the student population will be heterogeneous, which is challenging but also foreshadows the future working environments of most graduates. Relatedly, it offers opportunities for group assignments in multidisciplinary teams.

When the transition to new or extended programs in statistics and data science is discussed, much attention typically goes to curriculum guidelines, but less attention is paid to the admission process. However, every academic year, admission and examination boards experience the importance of an efficient and well-calibrated admission process. Setting the required prerequisite knowledge too low implies that too many students will fail and hence will not earn the degree after three to four years of study. This has to be avoided for many reasons, the most important of which is the future of the student concerned, next to cost-efficiency considerations. Setting the prerequisite knowledge too high would withhold the unique opportunity for realizing the dream to become a professional statistician/data scientist from those who have gaps in their knowledge but the potential to fill them in and make required progress en route.

The starting point is an unambiguous set of verifiable entry competencies, clear communication about it to candidate students, an admissions process built upon verifiable information, and valid proof that the candidate meets the entry criteria. In our view, an applied master's in data science program should be open not only to those with a bachelor's degree in statistics, mathematics, and engineering, for example, but also to bachelor's (and master's) degree holders in fields such as biology, life sciences, medicine, economics, chemistry, sociology, and psychology. More generally, bachelor's and master's degree holders from fields where data are the basis of knowledge, evidence-based decisions are made, and sufficient quantitative knowledge and a high level of abstract thinking are required can be considered for admission. Clearly, more so than ever, individual assessment before entry will be needed, which is a challenging but necessary endeavor.

Evidently, in the transition from a statistics to a data science program, the admissions process needs to transition along with it.

5. WHAT?

Hicks & Irizarry (2018) started their paper by raising the question, "What is missing in the current statistics curriculum?" This is a very natural question for statistics programs that intend to widen their scope toward data science. An important perspective on this question comes from the job market. Zheng (2017) argues that classically trained statisticians nowadays find themselves frustrated and disappointed when they realize that their competitors from computational sciences perform better in job interviews and hackathons involving heavy computational skills. More and more job announcements explicitly ask for data scientists rather than statisticians.

The previous paragraph started from the assumption that a data science program is built on an existing statistics program. However, there are other strategies: One could start from scratch or from another existing program, such as computer science, engineering, or mathematics. In this article, we only discuss how a data science curriculum can be added to a statistics program. It is important to keep in mind the important implication that graduates from such programs should be data scientists and statisticians simultaneously. This has consequences that are discussed later.

We now make a distinction between two types of statistics programs. On the one hand, a program may have a horizontal structure, which indicates that tracks in certain specializations are offered in a modular system. Students choose modules to orient their studies toward a specialization (e.g., biostatistics, business statistics, social statistics, bioinformatics). Thus, each student may compose their own individual program (obviously with some restrictions). On the other hand, in a vertical program, all specializations have many courses in common (e.g., in the first year), and the specialization-specific courses are mostly offered in the second year. This structure typically leaves only limited room for elective courses.

When a horizontal statistics program is changed into a statistics and data science program, the new data science courses are typically common for all students. This ensures that all students can graduate with data science qualifications. If the data science courses were only offered as a module, not all students would necessarily acquire the data science skills. For statistics programs that extend their scope in this way, the names of the tracks could then be changed, for example, from "biostatistics" to "biostatistics and data science," i.e., adding the data science label to all tracks.

In a vertical structure, new data science subjects can be added to the common part (mostly in the first year) and a new specialization in data science can be created. In this way, all students will have a firm basis in statistics and data science, but students in the data science specialization can deepen their data science knowledge and skills. In Belgium, the new Master of Statistics and Data Science programs at KU Leuven and Hasselt University are examples of a horizontal and vertical structured curriculum, respectively. Despite this fundamental difference, both programs share a few courses.

Before addressing what data science topics can be included in a curriculum, we want to state the obvious truth that adding more courses to a curriculum necessarily implies a reduction of content in the existing courses. Thus, when changing a traditional statistics program to a statistics and data science program, quite a few statistical topics need to be sacrificed, courses merged, and content pruned from redundant or overly elaborate topics. The search for these topics is perhaps as important as the selection of new data science topics to be added. The challenge of adding new and deleting existing content should go closely together with the challenges of applying innovative teaching methods, optimizing learning efficiency, and controlling course and study load. Blended

or hybrid learning combining online with traditional classroom methods (e.g., flipped classrooms), tailored to the learning activities within or across courses, enriches students with important skills such as self-learning, self-monitoring, and self-adjusting, which are key skills for lifelong learning.

Hicks & Irizarry (2018) argue that the missing topics in current statistics curricula are related to computing, connecting, and creating. Computing refers to programming skills and general knowledge about computing infrastructures and organization (e.g., parallel computing, distributed computing, databases). Data scientists should be able to connect their skills to subject matter questions. By "creating," they mean that the data scientist should take part in the creative scientific process and that he/she must formulate new substantive research questions. They argue that courses should be built around diverse case studies, include computing in almost every aspect of the course and minimizing the use of mathematical notation. They continue by recommending that all course activities should realistically mimic a data scientist's experience. This is in contrast with typical courses in a conventional statistics curriculum, in which students almost always get well-defined homework assignments that start from clean data sets that are provided to them in a format that can be directly read by software. A similar observation was made by Zheng (2017), who concludes that statistics students do not get enough experience with the entire data analysis cycle. He also suggests exposing students much more to realistic situations with messy data. A similar data-centric approach is also put forward by Hardin et al. (2015). From this literature (and the references therein), there seems to be a broad consensus in favor of curricula with many data-centric courses that integrate computing skills, algorithmic thinking, and, of course, statistical thinking.

Although we agree with many of these suggestions, we do have a few remarks. Programs that aim at broadening their scope from statistics to statistics and data science, either horizontally or vertically, do not necessarily need to redesign all their statistics courses to make them data-centric or emphasize computational aspects. We believe that there also must be data scientists who have the skills to read and understand the large body of statistical literature. This requires a good understanding of the traditional statistical theories and conventions because they still stand and they form a strong basis for many data analyses. If graduates from data science programs were no longer capable of reading and understanding the vast statistical literature, then this knowledge would be lost to them. Moreover, we see advantages in deliberately integrating course activities for which no computers may be used. Thinking and reflecting on problems and solutions are also important skills; the current generation of students is often distracted when sitting in front of a computer. Also, for the design of experiments and studies, computers are not required in all phases. Another motivation for not redesigning all courses is more of a practical nature: Sometimes courses are shared with other programs. For example, a programming course (e.g., Python) may be offered to student groups from different master's programs (e.g., engineering, mathematics, and physics). The organization of a separate course for only the statistics and data science students may require too many resources. In such cases, as a compromise solution, the homework or project assignments could be differentiated between the student groups, or a separate tutor for computer labs could be assigned for the statistics and data science students. These adaptations come at a lower cost and may still help in targeting these shared courses to a data science audience.

Coming back to the position of mathematics in the program, we refer to De Veaux et al. (2017), who recommend a curriculum of mathematical foundations and statistical modeling. In fact, we can refer to this as the required mathematical foundation of statistics (i.e., probability and statistical inference) on the one hand and statistical techniques on the other. At the graduate level, it is important to connect to the knowledge and skills already acquired during prior education and to build upon these to deepen and refine the knowledge and skill sets. The data scientist should possess a sufficiently large tool kit of statistical models and other techniques, together with a clear understanding of their uses, advantages, and limitations from a mathematical and computational

perspective. They should be able to examine the properties of newly proposed methodology, including by themselves. This can, but need not, be based on mathematical arguments; there is also room for simulations, for example. So, even profound theoretical concepts should not be taught in a context-free way, but have to be linked, as generically as possible, to the statistical and data science practice. This connects to the fact that the successful data scientist should have the habit of following up on new evolutions in any of the underpinning disciplines, not just to know and use whatever is new, but also to assess merits and pitfalls in a systematic, critical way.

Earlier, we discussed the suggestions to make all courses data-centric and computing-oriented, and to design all courses around case studies and make the students start from extracting or scraping messy data from databases or the Internet. Another curricular element that may be considered to reach more or less the same goal is a capstone project. For example, in each semester, one capstone project may be part of the curriculum. It may be organized as a full week, in which no other lectures are scheduled. To some extent, this can be compared with a hack week, which was also proposed as a model for data science education by Huppenkothen et al. (2018). Students may work in small groups on a large real-world case study, and the methods and skills required to solve the problems should come from the other courses taught in that same semester. Lecturers of those other courses may voluntarily enter the computer lab during this project week and may help or discuss issues with the students, or they may even do some just-in-time teaching when appropriate.

Whether or not all courses become data-centric and make students go through the whole data analysis cycle, we also believe that there should be better connections between the courses as compared with many traditional statistics programs today, in the sense that at least some of the fundamental data management and programming skills should return in many courses. When, for instance, SQL (Structured Query Language) is only used in the data management course, students may be unclear as to why they have to acquire knowledge about this language.

Thus far, we have focused mainly on adding more computational skills to the curriculum, as well as bringing and keeping the students closer to the data. With respect to the former, there seems to be a consensus that both R and Python have their place in a statistics and data science program. Many new programs offer a course about data management, which may cover topics related to databases, querying, data representations, and related topics. Data wrangling may be part of such a course, or it may be part of another course (e.g., basic programming in R or Python). Data wrangling or data cleaning is important to prepare students for dealing with the whole data analysis cycle. We already mentioned the importance of algorithmic thinking for data scientists. It is important to first decide whether a curriculum wants algorithmic skills or algorithmic thinking as a learning outcome. If it is the former, then a separate course, or sufficient attention to algorithms in a programming course, may suffice. However, when algorithmic thinking is the objective, it should be integrated in many courses. Combinations of both approaches are, of course, also welcome.

Another consensus topic is data visualization. When data come to us as simple matrices, most scientists know how to visually explore the data, and even for high-dimensional data, many new visualization tools exist. It becomes less obvious when data can no longer be represented as a matrix or when multiple data sources or data types need to be combined for visual exploration (e.g., networks, audio, video). Because of these complexities, a separate course on this topic is necessary, or it should be sufficiently covered in courses on data management, data visualization, and computing.

An important part of the curriculum should be devoted to awareness and active pursuit of good scientific practices, ethical behavior, and protection of privacy (Keller et al. 2016). The core curriculum should contain goals and methods for reproducible research, including version control, well-annotated software, seamless linking of data, analysis, and reporting. The use of packages and

programs such as `R markdown`, `Sweave`, Overleaf, and Jupyter notebooks, and familiarization with software development platforms such as GitHub, are typically part of courses on programming in R or Python and data management, but such tools should be used consistently in different learning units or courses across the whole study program.

Saltz et al. (2018) identified 12 key ethics themes that should be present in a data science curriculum. They conclude that no single existing code of conduct or ethics framework covers all issues and so there is a need for more research in this rather new area. They also make the point that apart from general ethical values that should hold for all data scientists, ethics highly depend on the context and the field of application. For example, working with sports data requires a different ethical framework than working with medical data. Although one may consider adding a course devoted to the ethics of data science, we believe that ethics should be touched upon in most of the other courses. In this way, the ethical considerations can be discussed in particular contexts (e.g., informed consent in a clinical trials course). In a similar fashion, legal aspects (e.g., privacy/the European Union General Data Protection Regulation) can be touched upon in courses when relevant. Indeed, the vast availability of so-called nondesigned data [Keller et al. (2016) estimate it to represent about 70% of the available data, made up of digital data and social media exchanges] requires careful reflection on privacy and corresponding measures to ensure confidentiality. It is interesting to see that time-honored techniques, such as multiple imputation (Rubin 1987), can be used to ensure confidentiality, as noted by Keller et al. (2016). Multiple imputation can preserve important structure in the data while avoiding the data being analyzed to reveal individual identities. Moreover, the statistical properties of such data, which are often poorly understood, should be given careful study—quite opposite to the common belief that the unbelievably vast amounts of data available obviate the need to worry about statistical pitfalls (e.g., bias).

The qualified data scientist should carefully understand the concept of differential privacy (see, e.g., Dwork & Roth 2014), in the sense that aggregate information (e.g., demographic information) or patterns in data (e.g., contact tracing in the context of an epidemic) can be made available, perhaps even publicly, while protecting the privacy of the individuals on which the data is based. In addition to understanding, the ethical need for it should be internalized, and the advantages but also pitfalls carefully understood. In other words, it should become a natural habit to think in terms of differential privacy whenever applicable.

Perhaps privacy may sound noncommittal, because at first it may seem immaterial to the core content, even to those who are teaching the courses. This concern may also hold for ethics and also for other competencies that, at first sight, are far from the core topic of a course. Even algorithmic thinking and computational skills may not sound evident to all lecturers who have to change their course from a statistics to a statistics and data science program.

It is thus important that every program starts from a well-formulated vision of the education of data scientists, and that this vision is communicated, understood, and supported by all lecturers. This vision should be translated into competencies and skills, which should be further made concrete as learning lines. It is the responsibility of the program committee to ensure a logical structure among the topics, arrange them into learning lines, and communicate to the lecturers what topics should be incorporated in what course. This is particularly important for ethics and legal aspects, as these usually do not belong to the core technical content of a course.

We have discussed ethical and legal aspects relevant to data science, but there are other skills and competencies that are equally important but not specific to data science. A good example is soft skills, such as communication and employability skills (e.g., stakeholder awareness and self-management). These skills should be incorporated into the curriculum in ways that will be relevant to data scientists.

Earlier we mentioned the role of a program committee as central to developing the vision for the program and the formulation and coordination of the learning lines to guarantee that all competencies are covered at the right place and time in the curriculum. When a traditional statistics program is transitioning to a statistics and data science program, it seems important to us that the program committee gets broadened as well, by attracting lecturers of the nonstatistical data science subjects (e.g., computer scientists).

Coming back to the relationship between data science and statistics, which is central to the development of a program, we like to stimulate thinking about the added value of statistical thinking and the statistical tradition.

In the development of a statistics and data science curriculum, one should also carefully think about connecting the knowledge, skills, and competencies the students have acquired in their undergraduate studies with the master's-level courses. For example, as indicated by De Veaux et al. (2017), careful attention should be devoted to data preparation and data management to deepen skills acquired at undergraduate level—or, if they have not been acquired yet, to mend this.

Models and techniques studied at undergraduate level should be taken to the next level. For example, models at undergraduate level are typically of a univariate or classical multivariate nature. However, there are many and potentially rich but complex correlated data structures that should be recognized and properly handled. In the same vein, a coherent treatment of missing data problems and solutions should be included. Especially with very big data collections, it is tempting to believe that there are enough data to overcome the missing data problem, which is, of course, a fallacy. The risk of bias resulting from incomplete data, confounding, and selective sampling mechanisms should be understood, and strategies should be offered to handle these well.

Unlike at undergraduate level, graduate education can benefit from a clear focus on one or a few substantive sciences. This need not be a uniform choice for a given curriculum, but there can be several tracks to cater to various interests and student subpopulations. Arguably, basic courses in the substantive science should be given, not with the aim to make the data scientist a specialist in the substantive area, but to give them the language, habits, and comfort to successfully interact with specialists in the area.

An interesting perspective is given by Hernán et al. (2019), who say that building a new data science curriculum should be taken as an opportunity to leave the conventional approach to statistical teaching behind, not only to include more computational and algorithmic competencies, but also to bring causal inference into the data science programs. This view may be considered too extreme, and a good balance should be struck across competencies.

6. WRAPPING UP AND LOOKING AHEAD

Many publications on the topic of teaching data science emphasize the need to structure the course activities to realistically mimic a data scientist's experience. We hold the opinion that several current applied master's of statistics programs offer such course activities and could thus be extended to a data science program, but we argue that there should be a good balance of different types of courses. Not all courses should be organized in this manner, as such activities often inevitably imply several concepts and knowledge to be fragmented and scattered and not discussed within a framework that offers sufficient background and insights. Therefore, there is, in our view, across parts of one single course, or preferably across courses, the need for applied experience courses as well as more focused courses on a particular topic, for example, a course on longitudinal data models or an advanced programming course. So, richness in types of courses and experiences for the student are in our view key to obtain a good balance in methodological insights, key competencies and skills, and so on.

The issues of ethics, privacy, and data protection and related regulation should receive sufficient attention in the curriculum. Passive knowledge and awareness of them is essential, but we highly recommend that students have active experience on these key issues, as part of a consulting class in a project course or within the framework of the master's thesis. Elliott et al. (2018) call attention to the importance of teaching cross-cultural ethics. Many statistics and data science study programs have a multicultural student population, including Western, Eastern, and African cultures. Cultural and also linguistic differences form barriers for ethical decision-making compatible across philosophical and theological models.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank the editors and the reviewers for their constructive comments and suggestions.

LITERATURE CITED

Cleveland WS. 2014. Data science: an action plan for expanding the technical areas of the field of statistics. *Stat. Anal. Data Min.* 7:414–17

Davenport TH, Patil DJ. 2012. Data scientist: the sexiest job of the 21st century. *Harv. Bus. Rev.* 90(10):70–76

De Veaux RD, Agarwal M, Averett M, Baumer B, Bray A, et al. 2017. Curriculum guidelines for undergraduate programs in data science. *Annu. Rev. Stat. Appl.* 4:15–30

Donoho D. 2017. 50 years of data science. *J. Comput. Graph. Stat.* 26(4):745–66

Dwork C, Roth A. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9:211–407

Elliott AC, Stokes L, Cao J. 2018. Teaching ethics in a statistics curriculum with a cross-cultural emphasis. *Am. Stat.* 72(4):359–67

Hallinen J. 2019. STEM education curriculum. *Encyclopedia Britannica*. **https://www.britannica.com/topic/STEM-education**

Hardin J, Hoerl R, Horton N, Nolan D, Baumer B, et al. 2015. Data science in statistics curricula: preparing students to "think with data." *Am. Stat.* 69(4):343–53

Heinemann B, Opel S, Budde L, Schulte C, Frischemeier D, et al. 2018. Drafting a data science curriculum for secondary schools. In *Koli Calling '18: Proceedings of the 18th Koli Calling International Conference on Computing Education Research*. **https://doi.org/10.1145/3279720.3279737**

Hernán MA, Hsu J, Healy B. 2019. A second chance to get causal inference right: a classification of data science tasks. *Chance* 32(1):42–49

Hicks SC, Irizarry RA. 2018. A guide to teaching data science. *Am. Stat.* 72(4):382–91

Horton NJ, Hardin JS, eds. 2015. *Am. Stat.* 69(4)

Huppenkothen D, Arendt A, Hogg DW, Ram K, VanderPlas J, Rokem A. 2018. Hack weeks as a model for data science education and collaboration. *PNAS* 115(36):8872–77

Kane MJ. 2014. Commentary: Cleveland's action plan and the development of data science over the last 12 years. *Stat. Anal. Data Min.* 7(6):423–24

Keller SA, Shipp S, Schroeder A. 2016. Does big data change the privacy landscape? A review of the issues. *Annu. Rev. Stat. Appl.* 3:161–80

Koomen MH, Rodriguez E, Hoffman A, Petersen C, Oberhauser K. 2018. Authentic science with citizen science and student-driven science fair projects. *Sci. Educ.* 102(3):593–644

Mikroyannidis A, Domingue J, Phethean C, Beeston G, Simperl E. 2018. Designing and delivering a curriculum for data science education across Europe. In *ICL 2017: Teaching and Learning in a Digital World*, ed. ME Auer, D Guralnick, I Simonics, pp. 540–50. New York: Springer

Pittard V. 2018. *The integration of data science in the primary and secondary curriculum*. Rep., R. Soc., London. **https://royalsociety.org/topics-policy/publications/2018/integration-data-science-in-primary-secondary-curriculum/**

Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley

Saltz JS, Dewar NI, Heckman R. 2018. Key concepts for a data science ethics curriculum. In *SIGCSE'18: Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pp. 952–57. New York: ACM

Wild CJ, Pfannkuch M. 1999. Statistical thinking in empirical enquiry. *Int. Stat. Rev.* 67:223–65

Yavuz FG, Ward MD. 2020. Fostering undergraduate data science. *Am. Stat.* 74:8–16

Zheng T. 2017. Teaching data science in a statistical curriculum: Can we teach more by teaching less? *J. Comput. Graph. Stat.* 26(4):772–74

Annual Review of Statistics and Its Application

Volume 8, 2021

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics